SRMv22 at RAL

John Gordon

Pre-GDB

6 February 2007

HSM at RAL

- RAL is replacing its home-grown HSM with Castor2
- This was successfully used with SRMv1.1for CMS CSA06 and is being rolled out to other experiments
- Shaun de Witt of RAL is the lead developer of SRMv1.1 and 2.2 for Castor so we feel confident we understand it.

- For SRMv1.1 Castor deploys a separate endpoint for each storage class

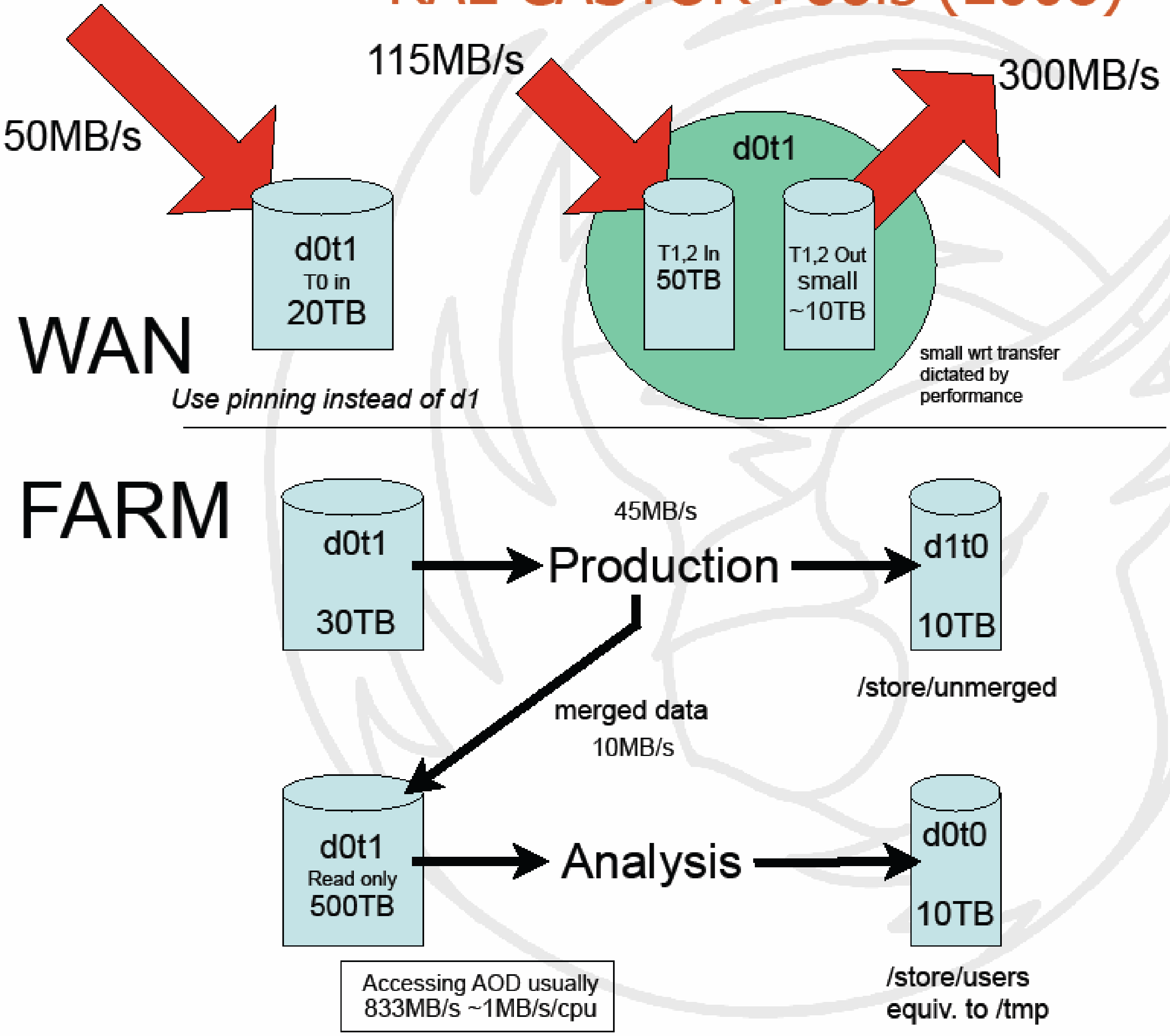
Outline

- Requirements
- Implementation
- Issues

Requirements

- We have had detailed discussions with CMS and feel that we have a clear idea what they need and want.
  - Need to clarify WAN and LAN differences
  - Stated no requirement for SRM from LAN
- We have started discussions with ATLAS and LHCb but do not yet understand to the same level of detail.
  - To be continued
- disk0tape1, disk1tape1, disk1tape0
- Multiple storage tokens per VO within a storage class

RAL CASTOR Pools (2008)



50MB/s

115MB/s

300MB/s

d0t1
T0 in
20TB

d0t1

T1,2 In
50TB

T1,2 Out
small
~10TB

small wrt transfer
dictated by
performance

WAN

Use pinning instead of d1

FARM

d0t1

30TB

45MB/s

Production

d1t0

10TB

/store/unmerged

merged data

10MB/s

d0t1
Read only
500TB

Analysis

d0t0

10TB

Accessing AOD usually
833MB/s ~1MB/s/cpu

/store/users
equiv. to /tmp

Data Types:

AOD - summary data. On disk always. Periodically in transfer to T1,2 at high rate.

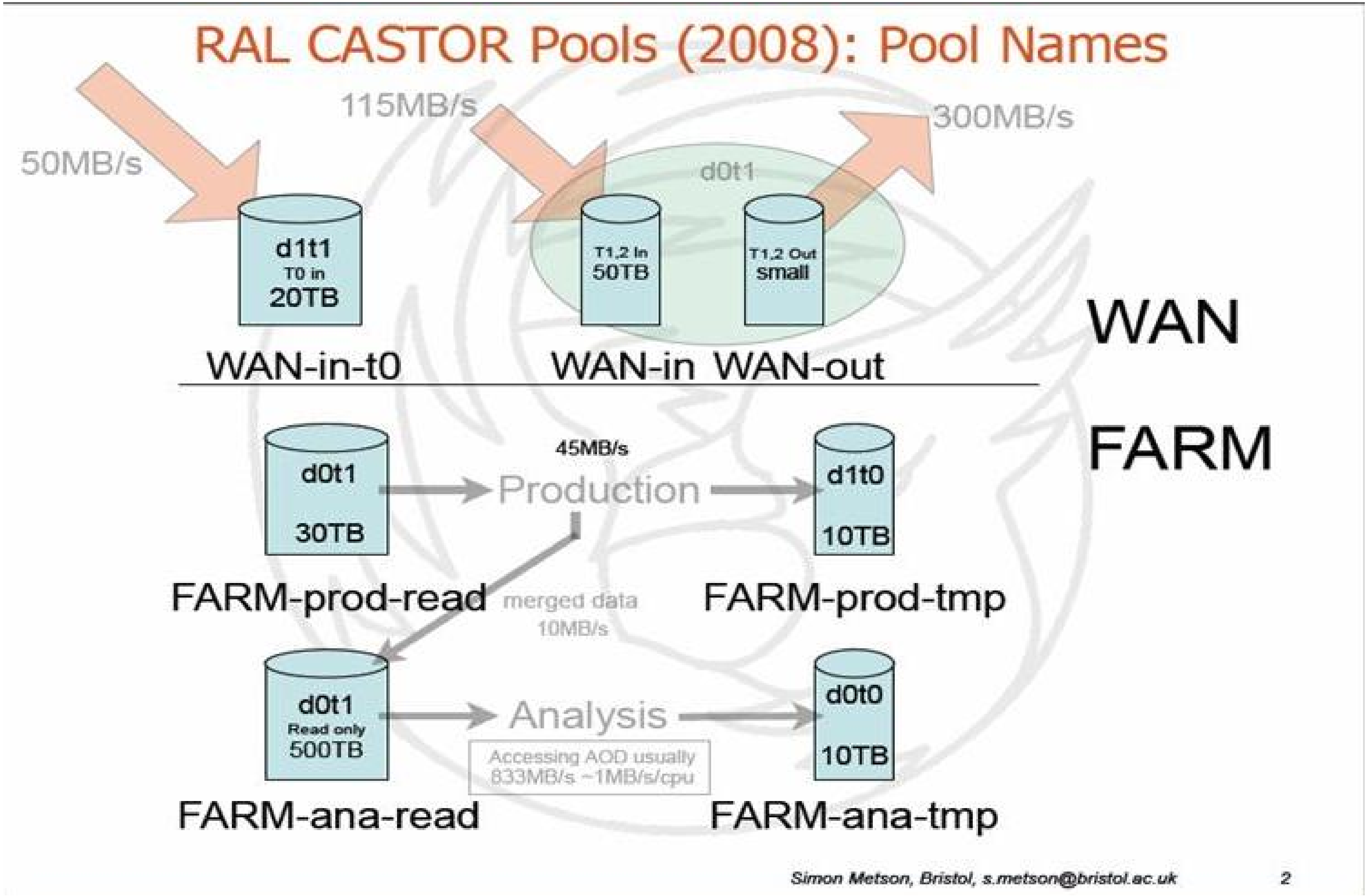
RECO - Produced at T0, reprocessed at T1's 3* per year. Should be on disk.

SIMRECO - Produced at T2's, should be on tape at least.

RAW - Data from detector, transferred from T0. On disk.

SIMRAW - Produced at T2's, should be on tape at least.

RECO, SIMRECO, RAW and SIMRAW are custodial data and must be stored on tape.



RAL CASTOR Pools (2008): Pool Names

50MB/s

115MB/s

300MB/s

d0t1

d1t1
T0 in
20TB

T1,2 In
50TB

T1,2 Out
small

WAN-in-t0

WAN-in WAN-out

WAN

FARM

45MB/s

d0t1

30TB

Production

d1t0

10TB

FARM-prod-read

merged data
10MB/s

FARM-prod-tmp

d0t1
Read only
500TB

Analysis

Accessing AOD usually
833MB/s ~1MB/s/cpu

d0t0

10TB

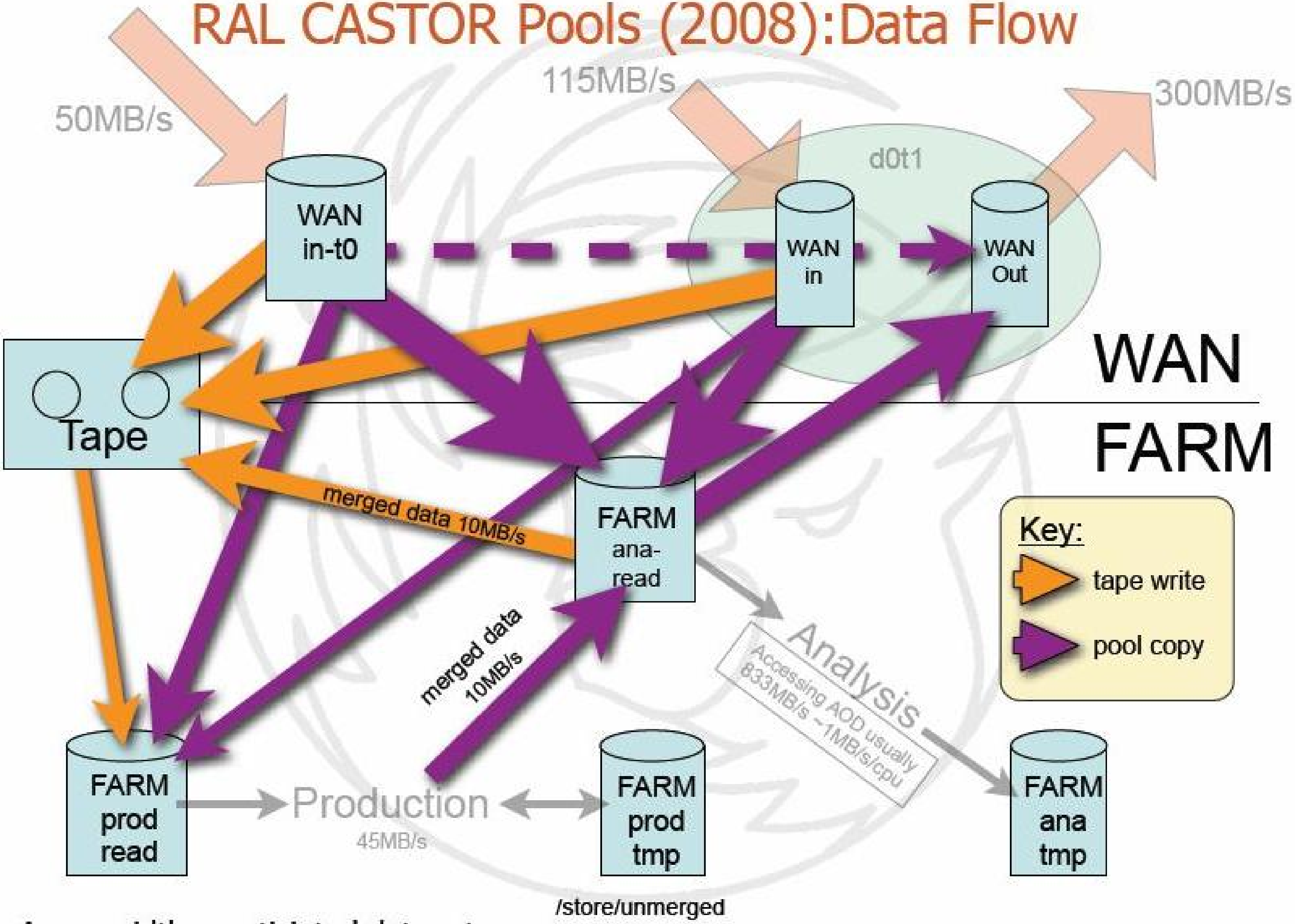
FARM-ana-read

FARM-ana-tmp

Simon Metson, Bristol, s.metson@bristol.ac.uk

2

RAL CASTOR Pools (2008):Data Flow

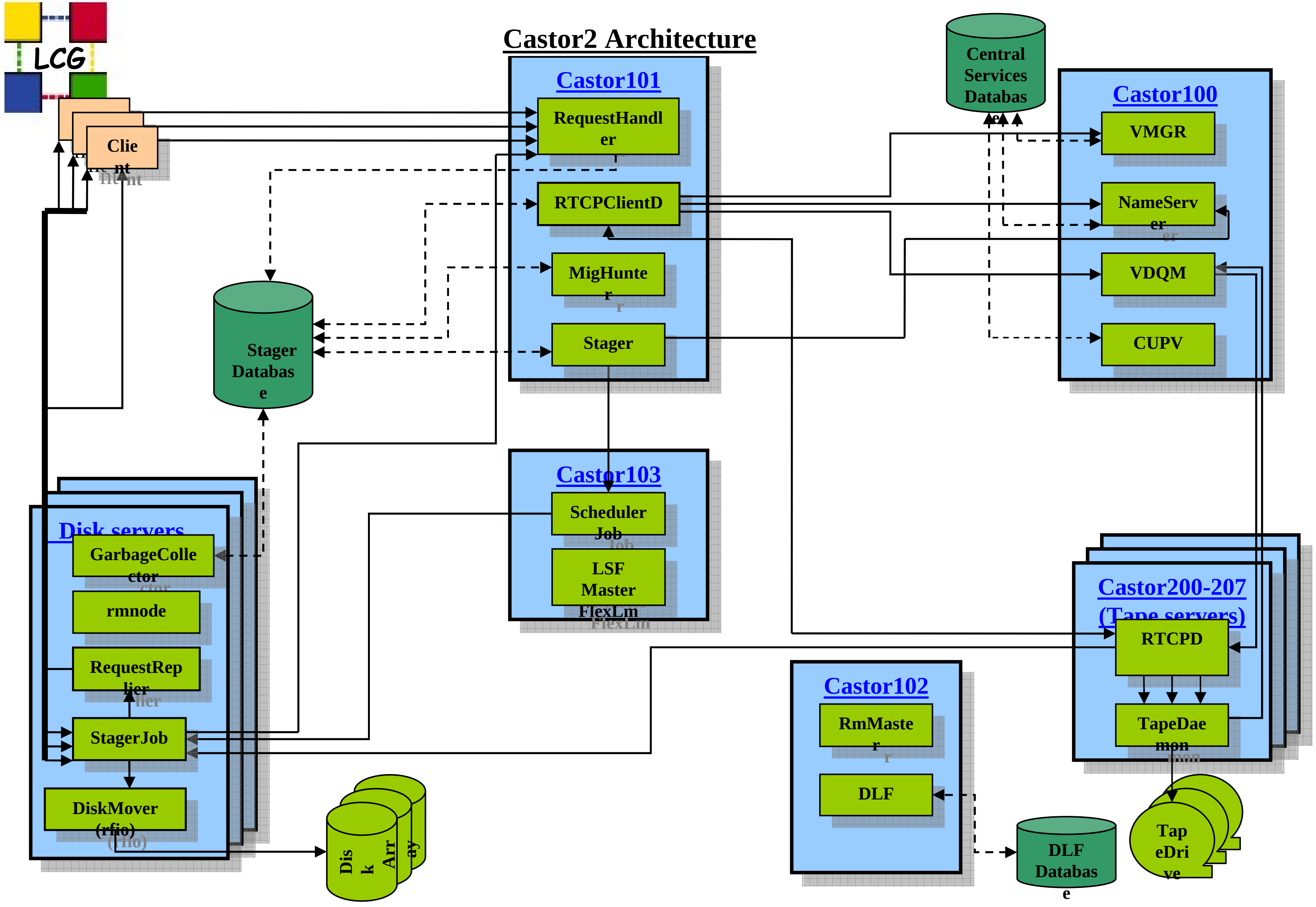


Arrow width ∝ anticipted data rate

Implementation – SRMv22

- SRM storage classes map onto Castor service classes at the file level so it is easy to implement these with a flat Castor file structure for all VOs.
- We won’t do this for various reasons
  - Fair shares of bandwidth between VOs.
  - VOs filling up servers affecting others
- We will give LHC experiments their own disk servers in a number of disk pools onto which we will map storage tokens
  - Different pools possible for different storage tokens with the same storage class
- Other smaller VOs may share a pool for everything

Castor2 Architecture



Castor101

RequestHandler

RTCPClientD

MigHunter

Stager

Central Services Database

Castor100

VMGR

NameServer

VDQM

CUPV

Client

Stager Database

Castor103

Scheduler Job

LSF Master FlexLm

Disk servers

GarbageCollector

rmnode

RequestReplier

StagerJob

DiskMover (rfio)

Disk Array

Castor200-207 (Tape servers)

RTCPD

TapeDaemon

TapeDrive

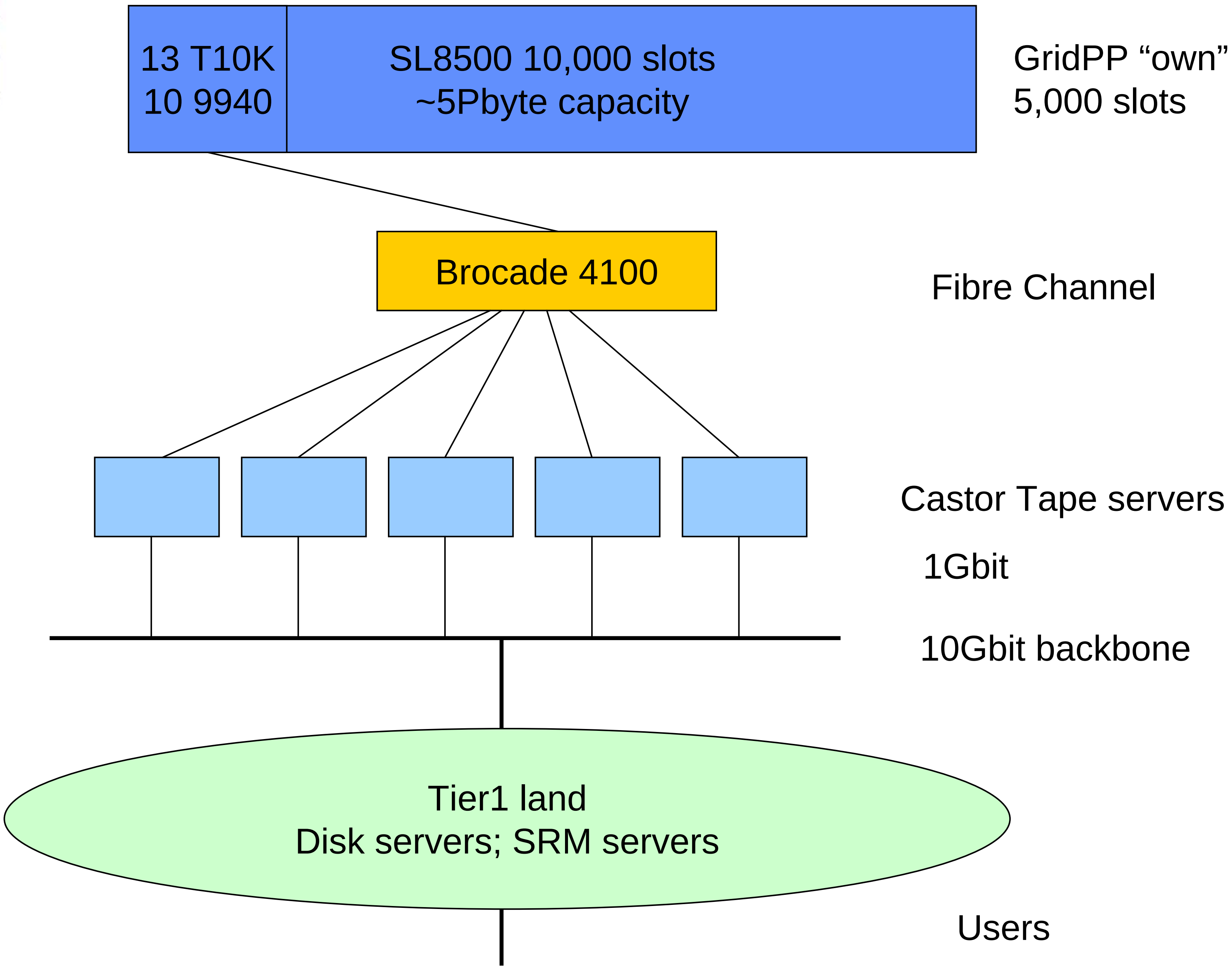
Castor102

RmMaster

DLF

DLF Database

j.c.gordon@rl.ac.uk



13 T10K
10 9940

SL8500 10,000 slots
~5Pbyte capacity

GridPP “own”
5,000 slots

Brocade 4100

Fibre Channel

Castor Tape servers

1Gbit

10Gbit backbone

Tier1 land
Disk servers; SRM servers

Users

Hardware (end 2007Q1)

By end of 2007Q1 we will have:

- Substantial expansion in disk capacity
  - 140 Disk servers - mainly Areca/3Ware with SATA
  - Providing 750TB of disk capacity
- 10000 slot SL8500 tape robot
  - 6 T10K drives dedicated to HEP/CASTOR
  - 6 9940 drives shared with other HEP VOs (dCache)
- 850TB media
  - 550TB on T10K
  - 300TB on 9940
- Additional drives and media planned in FY07 as understanding of CASTOR requirements grows
- Database architecture moving to RAC and data-guard for resilience and failover
- Separate Castor instance for Diamond and non-PP usage

Known Unknowns

- Castor2 support for disk1
  - But there is a plan
- Support for VOMS roles/groups
  - Current ACL model is uid/gid-based
  - Will LCMAPS configuration for Job priority also work?
- Performance
  - Export pools are small but require high bandwidth
    - Eg CMS T1, T2 out 300MB/s
  - May need special hardware or just spread across many servers
  - Share with other VOs to achieve high peaks,\low averages

Configuration Issues

- The interesting question is. For a VO –
- is it better to segment the storage and separate the flows into multiple pools
  - Stops interference
  - Allows specialist hardware if available
- Or run with a single big pool and average out all the I/O
  - Avoids small pools
  - allows more servers to be active at any time
- We don't know the answer to this but as CMS were keen to try a structured approach we will try it and see what we learn.